

基于分组残差及变换特征维度的目标检测网络

王改华^{1,2)}, 曹清程^{3)*}, 甘鑫³⁾, 翟乾宇³⁾

¹⁾ (天津科技大学人工智能学院 天津 300457)

²⁾ (北京中科慧眼科技有限公司 北京 100085)

³⁾ (湖北工业大学电气与电子工程学院 武汉 430068)

(102000235@hbut.edu.cn)

摘 要: 针对目标检测网络中低效的特征提取和重要特征难以聚焦的问题, 基于分组残差及变换特征维度, 提出一种改进的目标检测网络. 首先引入一个主干网络结构 G-ResNet50 作为特征提取网络, 以充分利用卷积的通道信息核; 然后提出广域感受野空间注意力机制, 强化重要特征的聚焦能力; 最后比较不同损失函数的回归效果, 并采用最佳的损失函数进一步提高网络检测目标的准确度. 在 PASCAL VOC 数据集和 COCO 数据集上的实验结果表明, 在相同的实验环境和设备下, 所提网络分别实现了 79.1% 的 mAP 和 35.7% 的 AP 值, 同时取得了 34.1 幅/s 的检测速度; 与 FCOS, ATSS, PAA 等网络对比, 该网络具有更好的目标检测效果.

关键词: 目标检测; 主干网络; 空间注意力机制; 损失函数

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20084

Object Detection Network Based on Grouped Residuals and Transformed Feature Dimensions

Wang Gaihua^{1,2)}, Cao Qingcheng^{3)*}, Gan Xin³⁾, and Zhai Qianyu³⁾

¹⁾ (College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin 300457)

²⁾ (Beijing Smarter Eye Technology Co., Ltd., Beijing 100085)

³⁾ (School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068)

Abstract: Aiming at the problem of inefficient feature extraction in the network and the difficulty of focusing important features, an improved object detection network is proposed based on grouping residuals and transforming feature dimensions. First, a backbone network structure G-ResNet50 is introduced as a feature extraction network; then a multiple receptive field spatial attention mechanism is proposed to strengthen the attention and focus ability of important features; finally, the regression effects of different loss functions are compared and the best loss function is adopted. Experimental results show that the proposed network can balance the relationship between calculation and detection accuracy, and capture more global information at the same time, which greatly improves the detection accuracy of the network. Compared with the comparison network, the network achieves better performance on the PASCAL VOC and COCO datasets.

Key words: object detection; backbone network; spatial attention mechanism; loss function

收稿日期: 2022-12-02; 修回日期: 2023-05-29. 基金项目: 国家重点研发计划(2017YFB1302400). 王改华(1979—), 女, 博士, 副教授, 硕士生导师, 主要研究方向为模式识别、深度学习; 曹清程(1998—), 男, 硕士研究生, 论文通信作者, 主要研究方向为深度学习、目标检测; 甘鑫(1996—), 男, 硕士研究生, 主要研究方向为深度学习、目标检测; 翟乾宇(1998—), 男, 硕士研究生, 主要研究方向为深度学习、图像语义分割.

目标检测是一种对图像或视频中的对象进行分类和定位的计算机视觉技术,广泛应用于行人检测^[1]、车辆检测^[2]、人脸检测^[3]、医学图像处理^[4]和遥感图像目标检测^[5]等方向。

基于深度学习的目标检测算法分为2类:两阶段和一阶段检测算法。通常,两阶段算法在精度上有优势,而一阶段算法在速度上表现更佳。

两阶段算法首先通过启发式方法或卷积神经网络(convolutional neural network, CNN)生成一系列稀疏候选框,然后对这些候选框进行分类和边界框回归。近年来,关于两阶段的目标检测算法层出不穷。2015年, Ren等^[6]提出 Faster R-CNN 算法,实现了一个区域建议网络(region proposal network, RPN)结构,显著地提高了目标检测速度;2018年, Cascade R-CNN^[7]通过级联检测网络实现不断优化效果,它的几种检测网络对不同交并比(intersection over union, IoU)阈值确定的正负样本都进行训练,达到了不错的效果;2019年, Pang等^[8]提出 Libra R-CNN,其中包含IoU平衡采样器、平衡特征金字塔和平衡 L_1 损失3个新颖的组件,这些组件可以减少样本、特征和目标层面的不平衡,提高了检测性能;2020年, Duan等^[9]提出 CPNDet,实现了一种无锚框两阶段框架,首先通过寻找潜在的角关键点组合提取多个目标提议,然后通过独立的分类阶段为每个提议分配一个类标签;2021年, CenterNetv2^[10]展示了从任意最先进的单级检测器构建概率两级检测器,同样提升了检测性能。上述两阶段算法具有较高的检测精度,然而,检测速度低于一阶段网络也是一个不争的事实。由于目前的目标检测任务更倾向于复杂的场景,因此对速度和效率的要求更加迫切。鉴于此,新提出的大多数目标检测网络越来越集中于一阶段算法。

一阶段算法的特点是在特征图上均匀地执行密集采样。2017年, Lin等^[11]提出 RetinaNet 方法来验证 Focal Loss 的有效性,解决了前景和背景不平衡的问题;2019年, FCOS^[12]通过逐像素预测的方式使得检测过程中不再需要 anchor box,避免了复杂的计算,显著地减少了所需的内存量;2020年, CentipetalNet^[13]使用向心移位配对来自同一实例的角关键点,预测角点的位置和向心偏移,并将其匹配到偏移结果对齐的角,提高了检测精度;2021年, Dynamic head^[14]展示一种新颖的动态头部框架,将目标检测头与注意力统一起来,在没有任何多余计算开销的情况下显著地提高了目标检测头的表达能力:这些一阶段算法使用不同的尺度和

纵横比进行采样,然后直接进行分类和回归,由于过程只需要一步,因此速度很快;然而,一阶段算法的准确度时常会低于两阶段算法。随着深度学习的不断发展和目标检测技术的不断优化,上述方法已经不能满足日益精细化的检测要求,如何使得一阶段网络在保持检测速度的同时也能达到较高的检测精度,成为当前工作的重点。

(1) 大多数网络使用的特征提取主干结构为 ResNet50 或 ResNet101,而这些仅仅是普通 1×1 和 3×3 卷积的交替叠加,随着网络深度的不断增加,深度卷积容易丢失语义信息,同时存在大量的特征损失,极大地影响了特征提取的能力。此外,随着网络深度的增加,模型计算量也会增加,网络往往没有充分利用卷积核的通道信息。为了解决这些问题,本文提出一种引入分组卷积的主干网络,更深入地组织每个卷积块之间的连接路径,丰富每个卷积之间的融合方法,有效地减少特征转移过程中的损失。

(2) 大多数网络仅限于传统的主干、颈部和检测头设计模式,没有考虑嵌入注意力机制后的增益效应。本文通过分析目标检测网络中主干、颈部和头部之间的连接关系,找到嵌入注意力机制的最佳方式,在不显著增加模型参数的情况下,极大地提高了网络对重要特征的注意力;本文网络并非仅仅将早期已提出的注意力机制直接添加到某个区域,而是对其进行改进和优化。已有的大多数注意力机制仅从单一维度设计,并且倾向于使用常见的方法,如三维的卷积和池化。三维特征的卷积、池化可以带来很好的效果,而二维特征应用并不广泛,其应用效果很难超越三维特征。然而,大多数工作并没有从二维和三维的深度细致结合角度展开,忽略了跨维度融合的特殊效果,因此专注于不同维度网络的交叉融合,从更广阔的角度改进注意力机制显得尤为重要。二维和三维的不断变换可以迫使特征信息不断地改变其呈现方式,使信息在缩放和扩展中更加活跃,也更有利于融合和交叉特征信息,提高其专注于特征图中关键区域的能力。

基于以上理论和方法,本文提出基于多通道分组残差与广域感受野空间注意力(multiple receptive field spatial attention, MSA)机制的目标检测网络。首先提出一种主干网络结构 G-ResNet50 (grouped residual ResNet50),使用分组残差结构提高主干网络提取特征信息的能力;然后提出 MSA 机制,从多个感受野中捕获像素级特征信息,通过特征维度的不断变换,增强重要的特征信息且抑制了不

重要的信息;最后将 G-ResNet50 和 MSA 机制作为即插即用的模块,添加到一些经典的目标检测网络中,如 FCOS, ATSS^[15]和 FoveaBox^[16],并在 PASCAL VOC 和 COCO 2017 数据集上进行实验,验证本文网络的有效性和普适性。

1 相关工作

1.1 主干网络结构

主干网络用于提取目标的不同尺度特征。早期,基于 CNN 的目标检测模型多使用 VGG^[17]作为主干网络,通过重复堆叠卷积和池化层提高特征提取和语义表达能力,但由于其层数只有 19 层,提取特征的能力有限;GoogLeNet^[18]的深度达到 22 层,宽度也得到了扩展;He 等^[19]提出一个跳跃连接的残差网络 ResNet,将浅层特征信息与后续层相结合;DenseNet^[20]通过特征重用和分流设置,极大地减少网络参数的数量并缓解了梯度消失问题;Darknet^[21]结合 ResNet 的特点,既保证特征的超强表达,又避免了更深的网络带来的梯度问题;ResNeSt^[22]展示一个模块化架构,将通道注意力应用于不同的网络分支,在捕获跨特征交互和学习不同表示等方面取得了良好的效果。

1.2 注意力机制

注意力机制广泛用于深度学习任务,如目标检测^[23-24]、语义分割^[25-28]和实例分割^[29],分为软注

意力机制和硬注意力机制。软注意力机制包括空间、通道和混合注意力机制,其中,空间域指图像中相应的空间变换,通道域直接将信息集中在全局通道中。SENet^[30]是一种称为“Squeeze-and Excitation”的通道注意力机制,分为 Squeeze 和 Excitation 这 2 个步骤;SKNet^[31]是一种“选择性内核网络”,通过选择适当的卷积内核执行注意力机制;CBAM^[32]是一个有效的注意力模块,可以从通道和空间中获取注意力图;DANet^[33]是一种双重注意力网络,可以自适应地集成局部特征和全局依赖关系;ASNet^[34]引入了密度注意力网络,同时提供不同级别的注意掩码密度:这些方法证明,注意力机制可以提高网络的特征表达能力。

2 本文网络

本文网络的结构如图 1 所示,包含 G-ResNet50,特征金字塔网络(feature pyramid network, FPN), Expand, Head 和 MSA 共 5 个模块。其中, G-ResNet50 包括卷积、池化和残差模块,残差模块中又包含 C_1, C_2, C_3, C_4 ; Expand 中的 2 个特征层通过下采样由 FPN 的上层特征得到;之后的 5 个输出可以检测大、中和小 3 种尺寸的物体,自下而上和自上而下的连接保持了多尺度特征图之间的注意力信息流;最后,输出被传递到检测头,子网结构包括用于分类和回归的分支。

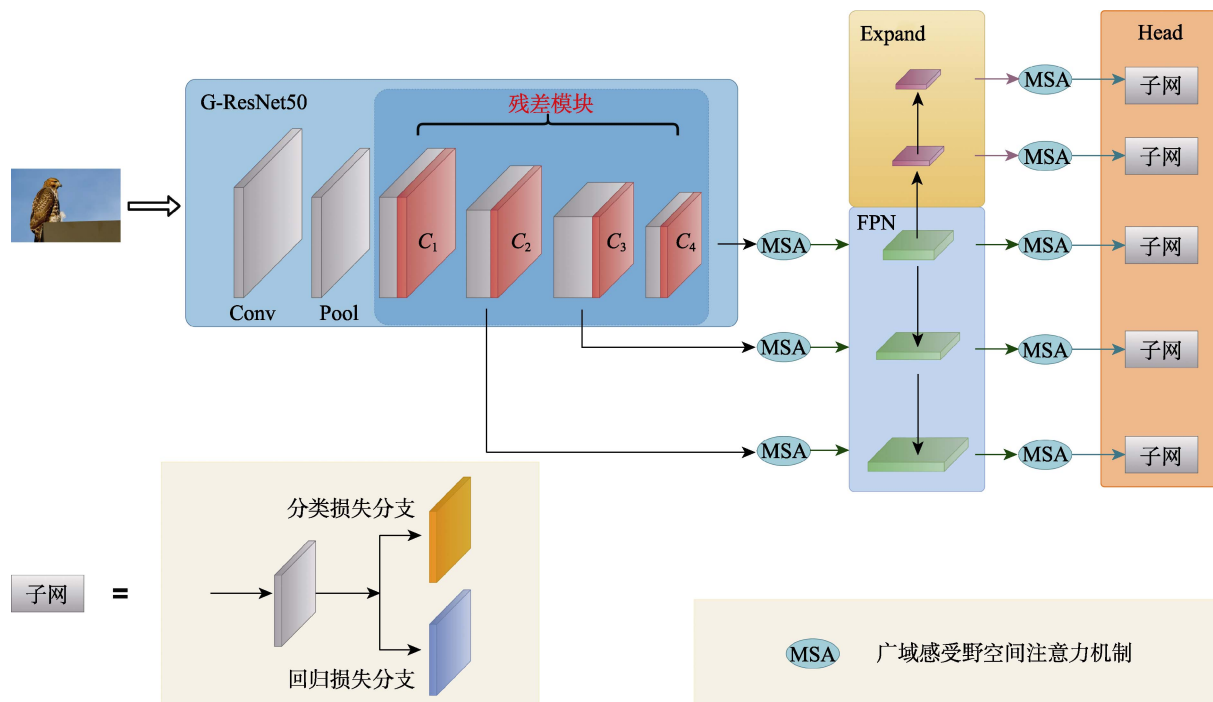


图 1 本文网络结构

2.1 G-ResNet50

原始 ResNet50 的结构如图 2 所示, 包含初始卷积、最大池化以及下面的 4 部分残差组. 其中, 每个残差组对应的特征图大小分别为 128×128 , 64×64 , 32×32 和 16×16 , 堆叠的数量分别为 3, 4, 6 和 5. 原始残差是残差组中的单个残差块结构, 包含

1 个 1×1 卷积、1 个 3×3 卷积、3 个 ReLU 激活函数和一个跳跃连接结构. ResNet50 的残差连接使得模型的深度不再是一个限制, 打破了神经网络的对称性, 提高了各层神经元的利用率; 然而, 深度的增加导致模型计算量也随之增加, 网络在增加深度的同时往往没有充分地利用卷积核的通道信息.

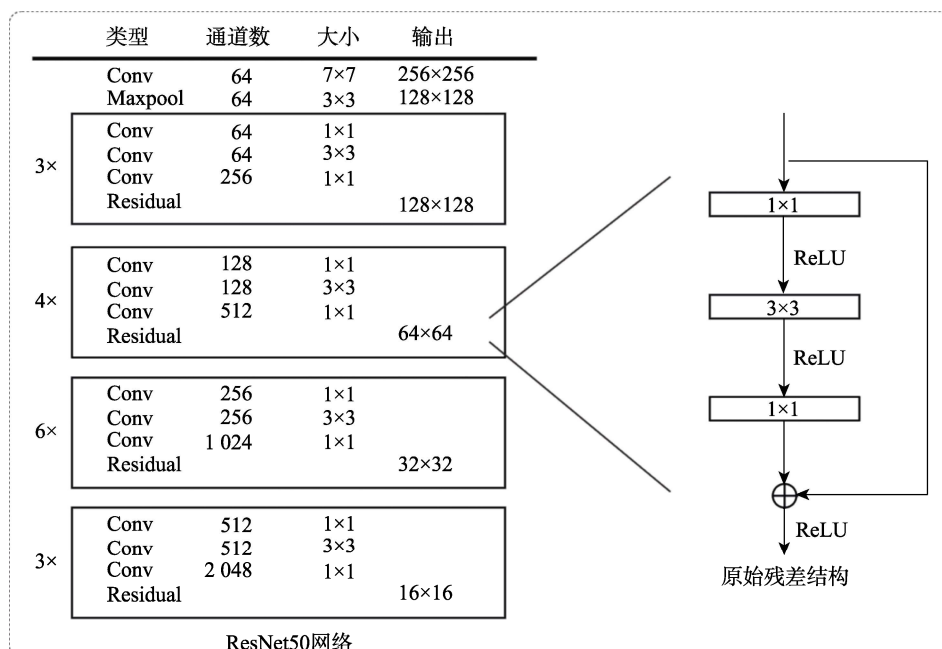


图 2 原始 ResNet50 结构

为了提高通道信息的特征提取能力, 本文提出一种分组残差结构, 取代原来的残差结构, 形成了一个增强的主干网络 G-ResNet50, 其结构如图 3 所示. (1) 特征信息首先通过卷积层和最大池化层. 与 ResNet50 相比, G-ResNet50 包含 8 部分残差组, 堆叠数量分别为 2, 1, 3, 1, 5, 1, 2 和 1, 其中包含 4 部分新残差结构. (2) 新残差结构中, 特征图首先经过 1×1 卷积, 然后进入分组卷积模块, 该模块将卷积层分为由 X_1 , X_2 , X_3 和 X_4 表示的 4 个通道组, 每组具有相同的宽度和高度, 通道数是输入特征图的 $1/4$, 其中, X_1 和 X_4 的特征信息分别与 X_2 和 X_3 交错生成 Y_2 和 Y_3 , 该操作可以显著提高信道特征信息的利用率; $Y_1 \sim Y_4$ 拼接融合后, 通过 1×1 卷积减少参数量; 外侧部分, X_1 和 X_4 经过 3×3 卷积增加通道维度, 然后与 1×1 得到的特征信息融合; 此外, 还使用 3×3 深度可分离卷积减少参数和计算成本, 随着网络层数的加深, 添加 1 个跳跃连接以避免梯度消失, 促进了梯度的反向传播并加速了训练过程; 最后, 通过 ReLU 激活函数输出特征图.

综上所述, 深化网络模型的过程中, G-ResNet50 在充分地利用卷积的通道信息核的同时, 避免特征信息的过度丢失, 控制了参数量和计算量, 与 ResNet50 相比, 具有更合理的结构和更好的性能.

2.2 MSA 机制

空间注意力机制是一种借鉴了人类视觉系统注意力机制的模型特征优化方法, 其原理是通过加权的方式增强图像上包含有效分类特征的区域, 并对包含无用特征的区域进行抑制, 防止模型学习到错误的分类信息. 针对特征图感受野和维度变换角度, 设计了 MSA 机制, 其结构如图 4 所示.

(1) 输入特征图 F , 设 $F \in \mathbb{R}^{C \times H \times W}$ 为输入张量, 其中, C , W 和 H 分别表示通道、高度和宽度. 为了减少参数的数量, 通过 3×3 的卷积将 F 的高和宽减半; F' 经过 3 个扩张卷积分支, 扩张率为 1, 2 和 3, 以及 1 个普通卷积. 在不降低感受野的情况下, 不同扩展率的空洞卷积可以减少空间特征的损失, 获得远程信息. F_1 , F_2 和 F_3 被重塑为 M_1 , M_2 和 M_3 , 它们具有相同的矩阵形状 $[H/2 \times W/2, C/2]$. 将 M_1 , M_2 和 M_3 相乘, 得到 3 个关系矩阵 N_1 , N_2 和 N_3 ,

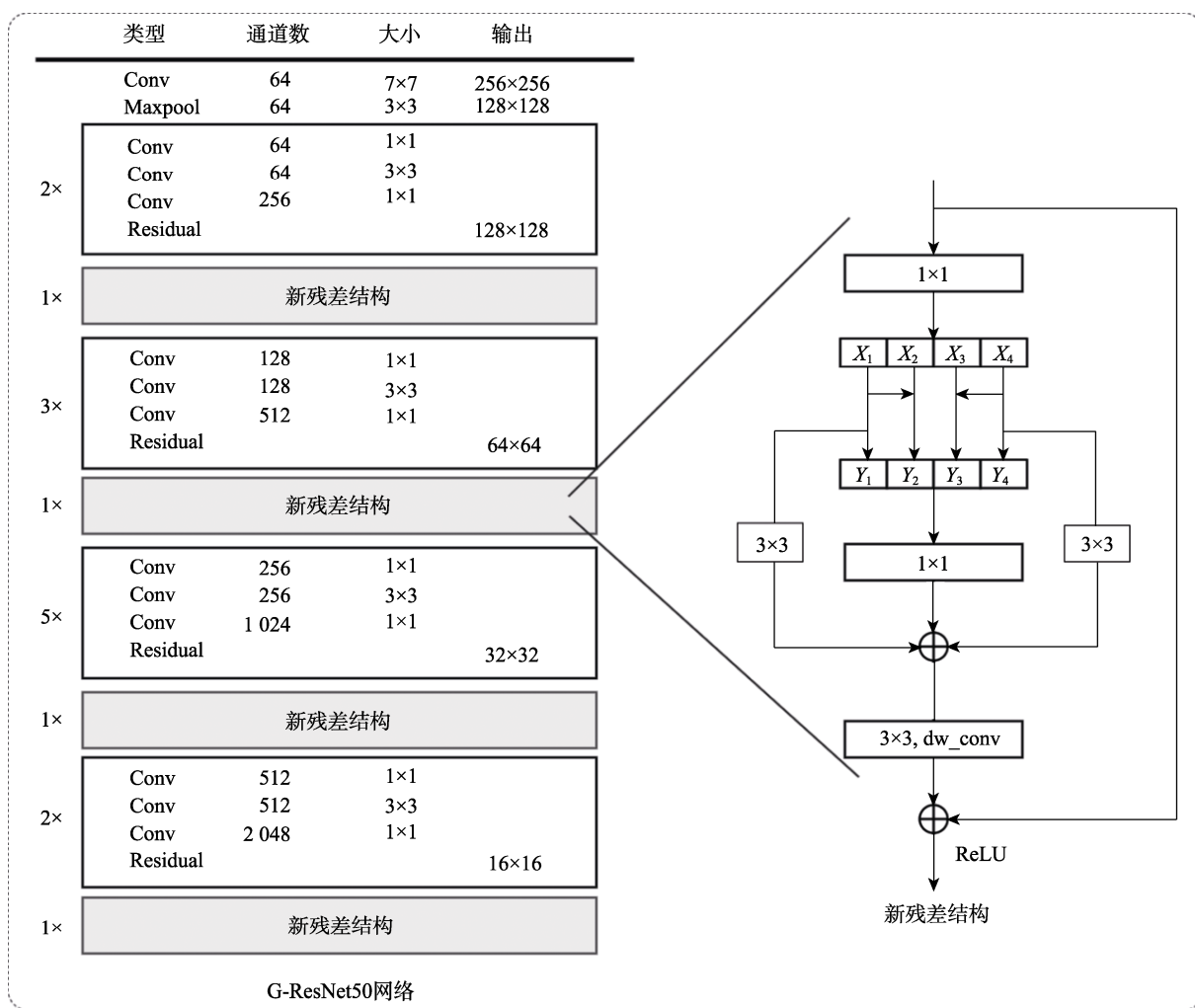


图 3 G-ResNet50 结构

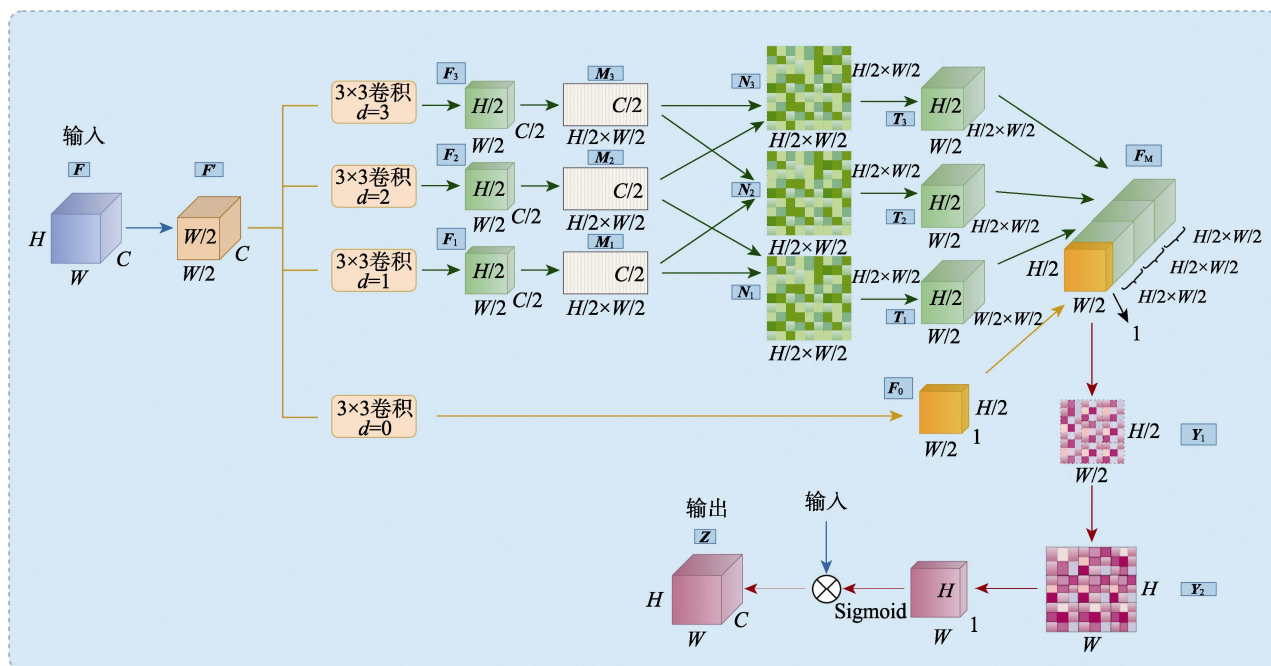


图 4 MSA 机制结构

关系矩阵中的每个值表示特征中每 2 个像素之间的关系. 值得注意的是, 3 个矩阵 $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ 及其转置矩阵可以通过多种方式相乘. 因此, 为了避免后期进行拼接时过多的计算量和参数, 减少部分矩阵之间的信息冗余, 从 $\mathbf{M}_1, \mathbf{M}_2$ 和 $\mathbf{M}_3$ 中选择 3 个具有代表性的交叉乘法方案, 得到 $\mathbf{N}_1, \mathbf{N}_2$ 和 $\mathbf{N}_3$, 保证了特征信息的相对完整性. $\mathbf{N}_1, \mathbf{N}_2$ 和 $\mathbf{N}_3$ 的计算公式为

$$\begin{cases} \mathbf{N}_1 = \mathbf{M}_2 \otimes \mathbf{M}_1^T \\ \mathbf{N}_2 = \mathbf{M}_1 \otimes \mathbf{M}_3^T \\ \mathbf{N}_3 = \mathbf{M}_3 \otimes \mathbf{M}_2^T \end{cases} \quad (1)$$

其中, $\otimes$ 表示矩阵乘法; $\mathbf{N}_1, \mathbf{N}_2$ 和 $\mathbf{N}_3$ 的形状分别为 $[H/2 \times W/2, H/2 \times W/2]$, $H/2 \times W/2$ 和 $H/2 \times W/2$ 分别为矩阵的行和列.

(2) 将 $\mathbf{N}_1, \mathbf{N}_2$ 和 $\mathbf{N}_3$ 重新整形为 $\mathbf{T}_1, \mathbf{T}_2$ 和 $\mathbf{T}_3$, 进行下一次特征融合操作, 其中, $\mathbf{T}_1, \mathbf{T}_2$ 和 $\mathbf{T}_3$ 的形状为 $[H/2 \times W/2, H/2, W/2]$.

对于另一个分支: (1) $\mathbf{F}'$ 通过一个 3×3 的普通卷积直接生成 $\mathbf{F}_0 \in \mathbb{R}^{1 \times H/2 \times W/2}$. 将 $\mathbf{F}_0$ 和 $\mathbf{T}_1, \mathbf{T}_2$ 和 $\mathbf{T}_3$ 组合, 得到特征 $\mathbf{F}^M \in \mathbb{R}^{(H/2 \times W/2) \times H/2 \times W/2}$. 该操作是 MSA 结构中的重要步骤, 将浅层信息, 尤其是边缘信息与经过多维变换和信息交叉后的深层信息融合, 提高了特征信息的表达能力. (2) $\mathbf{F}^M$ 被重塑为 $\mathbf{Y}_1$ 以产生注意力权重, 然后使用插值算法将注意力权重 $\mathbf{Y}_1$ 调整为 $\mathbf{Y}_2$, 得到与原始输入相同的高度和宽度; 再将 $\mathbf{Y}_2$ 整形为大小为 $[1, W, H]$ 的三维空间特征图, 经过 Sigmoid 函数与原始输入相乘, 得到最终输出

$$\mathbf{Z} = \text{mul}(\mathbf{F}, \sigma(f(\mathbf{Y}_2))) \quad (2)$$

其中, mul 表示乘法操作; σ 表示 Sigmoid 函数; $f(\cdot)$ 表示重塑操作.

与一般空间注意力相比, MSA 从多个感受野的角度捕获像素级的特征信息, 更重要的是, 其深入考虑了二维和三维变换带来的特殊效果. 连续的维度变换可以使得特征信息更加活跃和均匀地分布, 此时, 特征融合可以获得更丰富的信息. 最后, 在没有过多增加参数和计算量的情况下, 特征提取的效果得到了明显提升.

2.3 损失函数

检测头结构包括分类和回归分支, 负责锚标签的预测和位置的回归. 本文模型基于 VFNet^[35], 包括方差损失和广义 IoU(generalized IoU, GIoU)

损失^[36], 为了获得更好的检测效果, 使用 CIoU 损失^[37]代替 GIoU 损失进行位置回归. CIoU 考虑了重叠面积、中心点距离和纵横比 3 个重要的几何因素, 公式为

$$\text{CIoU} = \frac{\rho^2(a, b)}{c^2} + \alpha \nu \quad (3)$$

其中, a 和 b 表示 A 和 B 的中心点; $\rho(\cdot)$ 表示欧几里得度量; c 表示覆盖 2 个框最小包围框的对角线长度; α 表示正权衡参数, 定义为

$$\alpha = \frac{\nu}{(1 - \text{IoU}) + \nu} \quad (4)$$

重叠区域因子在回归中具有更高的优先级, 特别是对于非重叠情况; ν 表示衡量纵横比一致性的参数, 定义为

$$\nu = \frac{4}{\pi^2} \left(\arctan \frac{w^{\text{gt}}}{h^{\text{gt}}} - \arctan \frac{w}{h} \right)^2 \quad (5)$$

其中, w^{gt} 和 h^{gt} 表示真实边界框的宽度和高度, w 和 h 表示预测边界框的宽度和高度. 损失函数可以定义为

$$L_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(a, b)}{c^2} + \alpha \nu \quad (6)$$

3 实验结果及分析

为了得到更好的对比数据, 本文选择 COCO 数据集和 PASCAL VOC 数据集进行训练.

(1) PASCAL VOC 数据集. 目前, VOC 2007 数据集和 VOC 2012 数据集常用于目标检测任务. PASCAL VOC 2007 和 PASCAL VOC 2012 数据集分为车辆、家庭、动物和人 4 类, 共有 20 个子类别(加上背景 21 个类别). 由于 2 个数据集互斥, 本文使用 VOC 2007+ VOC 2012(train+val)进行训练, 同时使用 VOC 2007(test)进行测试.

(2) COCO 数据集. 该数据集是一个大型、丰富的目标检测、分割和字幕数据集, 以场景理解为目标, 从复杂的日常场景中截取获得, 并通过精确分割对图像中的对象进行标定. MS COCO 数据集图像分为训练集、验证集和测试集. COCO 数据集使用 AMT(Amazon mechanical turk), 通过在 Flickr 上搜索 80 个目标类别和各种场景类型来收集图像, 其包含 91 类对象, 328 000 幅图像和 2 500 000 个标签.

(3) 实验环境. 建立以 PyTorch1.6.0, CUDA10.0, CUDNN7.4 为深度学习框架的 Python 编译环境.

使用 MMDetection2.6 平台进行编译, 其由商汤科技(2018 COCO 目标检测挑战赛冠军)和香港中文大学联合开发的基于 Pytorch 的深度学习目标检测工具箱, 提供了多种经典的目标检测算法, 包括两阶段和一阶段算法, 也正因如此, 使得代码的使用和修改更方便, 实验数据也更有说服力。

(4) 实验平台. Intel Core i7-10700@2.90GHz CPU, 16GB 内存, NVIDIA GTX 2060 super 显卡, 500GB 硬盘. 输入图像统一调整为 512×512 像素; 数据集使用随机翻转、随机裁剪和填充操作; SGD 用于训练模型的损失函数; 设训练 batch size 为 4, 迭代周期为 24 epochs, 学习率为 0.001.

(5) 评价指标. 在 PASCAL VOC 数据集上, 选用目标检测中常用的平均准确率(mean average precision, mAP)指标进行评估; 在 COCO 数据集上, 除 AP 指标外, 同时选用 AP₅₀, AP₇₅, AP_S, AP_M 和 AP_L 作为评价标准, 其中, AP₅₀ 和 AP₇₅ 是取 IoU 阈值大于 0.50 和大于 0.75 的检测器的检测结果, AP_S, AP_M 和 AP_L 分别对应小、中、大型目标的检测准确度。

3.1 消融实验

为了评估本文网络, 使用不同的目标检测架构在各种任务中对 PASCAL VOC 和 COCO 2017 进行消融实验, 包括 FCOS^[12], ATSS^[15], FoveaBox^[16]和 VFNet^[35]. 该实验分为主干网络 G-ResNet50 和 MSA 机制 2 部分和注意力。

3.1.1 G-ResNet50 消融实验

将 ResNet50 替换为 G-ResNet50, 并保持所有其他参数一致, 在 PASCAL VOC 和 COCO 数据集上进行消融实验。

在 PASCAL VOC 数据集上的消融实验结果如表 1 所示, 其中, Res 表示使用 ResNet50 作为特征提取网络, G-Res 表示使用 G-ResNet50 作为特征提取网络. 可以看出, VFNet 的检测准确率最高, mAP 提高 2.2 个百分点. 在 20 个类别中具有更好的准确性; ATSS 网络达到 76.4% 的 mAP, 说明主干网络替换成 G-ResNet50 后, 比原来的 ResNet50 网络提高 1.4 个百分点; FoveaBox 和 FCOS 网络分别提高到 72.2% 和 74.4%, 同样取得了很好的检测效果; 与 ResNet50 相比, G-ResNet50 将标准卷积划分为更小的卷积组, 并融合不同的通道特征, 有效地提高了特征提取能力. 值得注意的是, 个别类别的检测准确率出现下降的情况, 如 bottle, bicycle, “盆栽”和“沙发”, 这是因为这些类别在训练实例中相对较少, 并且有些在图像中存在遮挡或者残缺, 并不完整。

表 1 PASCAL VOC 数据集上 4 种网络检测 mAP 对比 %

类别	FCOS ^[12]		FoveaBox ^[16]		ATSS ^[15]		VFNet ^[35]	
	Res	G-Res	Res	G-Res	Res	G-Res	Res	G-Res
aeroplane	71.3	74.9	80.5	82.4	78.9	81.5	80.7	81.5
bicycle	76.2	75.6	81.3	81.3	81.0	80.7	81.9	82.3
bird	77.5	76.2	76.0	74.9	74.7	73.3	77.6	81.0
boat	59.6	65.2	66.5	65.8	65.0	65.6	66.8	72.3
bottle	58.0	55.3	58.6	65.2	58.1	57.7	59.8	64.7
bus	80.7	80.0	81.5	81.5	79.7	84.0	83.3	84.5
car	82.0	82.1	84.9	85.4	84.8	85.7	85.9	86.2
cat	86.8	84.9	85.2	87.9	84.3	82.1	84.3	88.5
chair	55.2	56.2	58.9	58.1	58.7	60.2	60.2	61.3
cow	74.8	75.4	72.7	82.1	71.2	73.1	78.9	83.9
diningtable	62.4	63.4	67.7	69.9	68.5	70.7	68.0	69.5
dog	80.4	82.8	82.3	82.9	76.8	77.4	82.6	87.6
horse	70.9	73.8	79.5	83.6	81.6	81.7	84.7	85.8
motorbike	72.3	74.9	81.7	82.2	78.3	80.3	80.0	82.5
person	78.9	78.7	82.2	82.5	82.1	82.1	81.9	82.2
pottedplant	50.3	48.8	53.4	53.5	50.5	50.9	51.2	50.7
sheep	71.5	72.0	74.5	80.2	74.9	74.4	76.5	81.9
sofa	65.3	67.7	72.2	70.7	73.0	71.3	74.8	74.1
train	77.8	80.4	81.1	84.0	81.2	81.2	83.6	85.6
tvmonitor	75.8	76.1	73.6	73.7	75.0	73.7	77.2	77.8
mAP	71.4	72.2	74.7	76.4	73.9	74.4	76.0	78.2

注. 粗体表示最优值; Res 表示使用 ResNet50 作为主干网络, G-Res 表示使用 G-ResNet50 作为主干网络。

在 COCO 数据集上的消融实验结果如表 2 所示. 与 PASCAL VOC 数据集不同, COCO 数据集中图像海量, 种类繁多, 显著地增加了目标检测的难度, 很多网络的 AP 值只有 PASCAL VOC 数据集中 mAP 值的 1/2 甚至 1/3. 为了让实验更有说服力, 仍旧选择 FCOS, ATSS 和 FoveaBox 作为对比网络. 可以看出, ATSS 取得了最明显的改进, AP 值提高 0.9 个百分点, 特别是 AP_L 提高 1.9 个百分点; FCOS 的 AP 提高 0.7 个百分点, 同样满足了提升精度的要求; FoveaBox 的效果相对较差一些, AP 值仅提高了 0.3 个百分点, 这与 FoveaBox 的正负样本训练策略和损失函数有关; VFNet 的 AP 值从 34.1% 提高了 0.7 个百分点, AP_S 从 14.4% 提高到 16.2%; 在已具有如此高 AP 值的检测网络上, 本文网络仍然可以提高准确率, 充分证明了主干网络 G-ResNet50 结构的有效性。

3.1.2 MSA 机制消融实验

在 2 个数据集上, 将本文提出的 MSA 机制添加到 FPN 结构中进行消融实验。

在 PASCAL VOC 数据集上, 添加 MSA 机制前

表 2 在 COCO 数据集上 4 种网络的检测结果对比 %

网络	主干网络	AP	AP ₅₀	AP ₇₅	AP _s	AP _M	AP _L
FCOS ^[12]	Res	28.7	45.0	29.9	9.6	31.9	44.3
	G-Res	29.4	45.4	30.8	10.4	32.8	45.3
ATSS ^[15]	Res	32.1	48.7	34.1	13.2	35.3	50.1
	G-Res	33.0	47.9	35.3	15.1	36.4	47.1
FoveaBox ^[16]	Res	28.5	46.4	29.9	10.3	32.2	43.8
	G-Res	28.8	46.4	30.3	12.7	32.0	41.9
VFNet ^[35]	Res	34.1	49.7	36.5	14.4	37.6	50.5
	G-Res	34.8	50.3	37.3	16.2	38.0	49.8

注. 粗体表示最优值; Res 表示使用 ResNet50 作为主干网络, G-Res 表示使用 G-ResNet50 作为主干网络.

后的检测结果如表 3 所示. 可以看出, 加入 MSA 机制后, 每个网络的 mAP 提高了 1.3~3.1 个百分点; VFNet 的 mAP 为 79.1%, 且 20 个类别的检测准确率都明显高于原网络; ATSS 同样取得了优异的检测结果, mAP 为 76.1%; FCOS 和 FoveaBox 的 AP 值分别提高 1.7 个百分点和 1.3 个百分点, 也完成了提高检测精度的任务. 实验结果表明, MSA 机

表 3 添加 MSA 机制前后在 PASCAL VOC 数据集上的检测 mAP 对比 %

类别	FCOS ^[12]		FoveaBox ^[16]		ATSS ^[15]		VFNet ^[35]	
	×	√	×	√	×	√	×	√
aeroplane	71.3	75.5	80.5	76.3	78.9	82.1	80.7	82.3
bicycle	76.2	81.0	81.3	82.4	81.0	82.2	81.9	84.1
bird	77.5	78.6	76.0	81.3	74.7	78.3	77.6	82.1
boat	59.6	65.1	66.5	68.0	65.0	64.6	66.8	72.6
bottle	58.0	54.7	58.6	64.1	58.1	60.6	59.8	64.1
bus	80.7	81.9	81.5	81.4	79.7	82.8	83.3	84.6
car	82.0	83.7	84.9	84.0	84.8	84.9	85.9	85.9
cat	86.8	85.6	85.2	88.9	84.3	87.1	84.3	88.4
chair	55.2	56.4	58.9	59.6	58.7	60.6	60.2	62.0
cow	74.8	75.3	72.7	82.0	71.2	76.7	78.9	85.5
diningtable	62.4	64.7	67.7	67.6	68.5	69.9	68.0	72.3
dog	80.4	84.6	82.3	87.8	76.8	82.3	82.6	87.0
horse	70.9	79.6	79.5	80.0	81.6	82.7	84.7	86.7
motorbike	72.3	75.8	81.7	78.2	78.3	81.1	80.0	83.5
person	78.9	78.7	82.2	81.1	82.1	82.6	81.9	82.8
pottedplant	50.3	45.9	53.4	52.8	50.5	51.2	51.2	51.7
sheep	71.5	75.4	74.5	76.4	74.9	77.4	76.5	85.0
sofa	65.3	65.7	72.2	68.5	73.0	74.7	74.8	76.4
train	77.8	78.3	81.1	81.3	81.2	83.9	83.6	87.0
tvmonitor	75.8	75.1	73.6	78.4	75.0	76.6	77.2	78.0
mAP	71.4	73.1	74.7	76.0	73.9	76.1	76.0	79.1

注. 粗体表示最优值; 添加 MSA 机制表示为“√”, 未添加 MSA 机制表示为“×”.

制整合了不同感受野的特征图, 这些特征图在转换为二维矩阵后交错融合, 增强重要特征的关注度并减少了信息的冗余, 因此很好地提高了 AP 值.

在 COCO 2017 数据集上, 添加 MSA 机制前后的检测结果如表 4 所示. 可以看出, 每个网络的提高了 0.7~0.9 个百分点. 值得说明的是, 由于 COCO 2017 数据集通常包含许多复杂的对象, 少部分物体的类型、尺度和姿势往往是不确定的, 因此也会给检测带来一定的困难. 如加入 MSA 机制后, ATSS 和 VFNet 的大目标检测效果比原网络略差, 当然这可以认为是可接受范围内的实验误差, 并不影响 MSA 的有效性.

表 4 添加 MSA 机制前后在 COCO 数据集上的检测结果对比 %

网络	MSA	AP	AP ₅₀	AP ₇₅	AP _s	AP _M	AP _L
FCOS ^[12]	×	28.7	45.0	29.9	9.6	31.9	44.3
	√	29.6	46.5	31.0	10.8	33.1	45.4
ATSS ^[15]	×	32.1	48.7	34.1	13.2	35.3	50.1
	√	32.8	48.8	35.1	15.3	36.7	46.4
FoveaBox ^[16]	×	28.5	46.4	29.9	10.3	32.2	43.8
	√	29.2	47.9	30.4	11.6	33.0	44.5
VFNet ^[35]	×	34.1	49.7	36.5	14.4	37.6	50.5
	√	34.8	50.6	37.4	16.3	38.3	49.8

注. 添加 MSA 机制表示为“√”, 未添加 MSA 机制表示为“×”.

为了进一步验证 MSA 机制的作用效果, 将其与一些经典的注意力网络进行比较, 包括 SENet^[30], CBAM^[32], GSoP-Net^[38]和 ECANet^[39]. 在 PASCAL VOC 数据集上进行实验, 评价指标为 mAP, 实验结果如表 5 所示, 其中, 原始网络表示未加任何注意力机制. 可以看出, 本文提出的 MSA 机制的 mAP 为 79.1%, 高于其他注意力网络; 最重要的是, 与原始网络相比, 添加 MSA 机制后的检测效果得到了明显的改善; 本文网络结合多感受野卷积组, 极大地提高了特征图的感受野面积, 同时,

表 5 不同注意力网络对模型检测的 mAP 对比 %

注意力网络	mAP
原始网络	76.0
SENet ^[30]	78.3
CBAM ^[32]	78.5
GSoP-Net ^[38]	78.2
ECANet ^[39]	78.7
MSA(本文)	79.1

注. 粗体表示最优值.

二维和三维特征图之间的转换优化了特征融合的方法;此外,还将本文网络的复杂度控制在合理范围内,使得 MSA 机制的参数在略高于未添加时参数的前提下,具备了更好的性能.因此,实验结果充分证明了本文 MSA 机制的有效性.

3.2 损失函数对比实验

在 PASCAL VOC 数据集上,对不同损失函数的检测效果进行实验,结果如表 6 所示.

图 5 所示为 4 种回归损失函数的曲线对比.可以看出,CIoU 损失的回归曲线比较平滑;IoU 损失仅在预测框与真实框重叠、收敛较慢且初始损失较高

时才有效;DIoU 和 CIoU 损失可以比 IoU 和 GIoU 损失更快地收敛,而 CIoU 损失考虑了完整的几何因素,因此收敛速度和回归精度优于其他损失函数.

表 6 不同损失函数对模型检测的 mAP 对比 %

损失函数	mAP
IoU	75.7
GIoU	75.7
DIoU	75.6
CIoU(本文)	75.9

注. 粗体表示最优值.

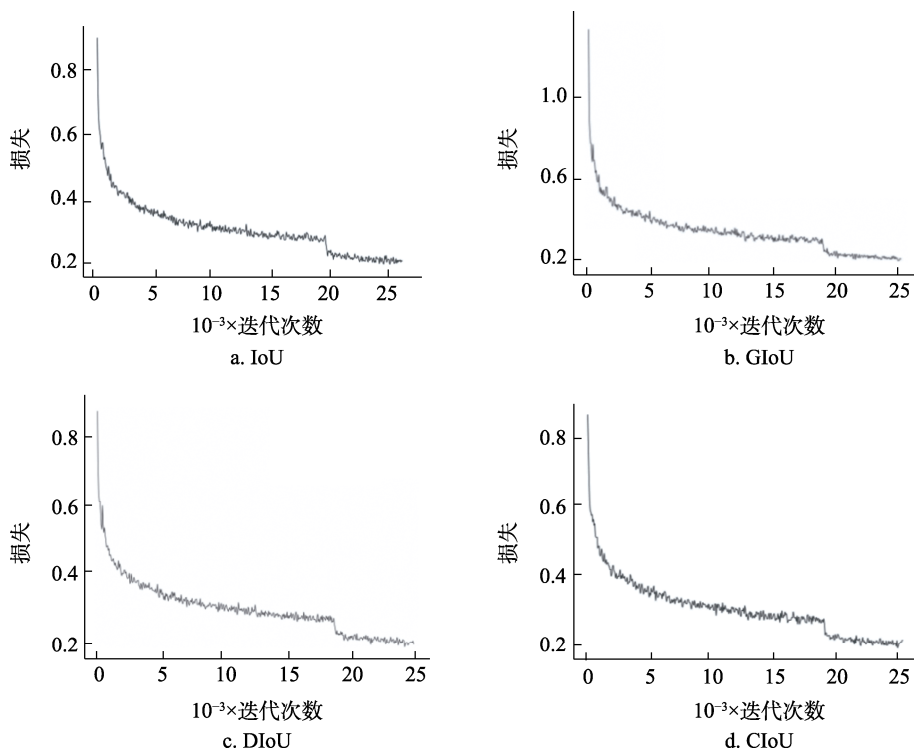


图 5 4 种回归损失函数的绘制曲线

3.3 对比实验

在 COCO 数据集上进行不断增加变量的实验,结果如表 7 所示.可以看出,虽然每个子模块有不同的输出结果,但它们都比不加任何附加模块的原始 VFNet(ResNet50+FPN)表现更好;原始输出的 AP 值为 34.1%,而用 G-ResNet50 替换 ResNet50 后,AP 值提高 0.6 个百分点,提高了特征提取能力;使用 MSA 机制后,AP 值则提高 0.7 个百分点,增强了重要特征的关注意力;同时添加 G-ResNet50 和 MSA 机制后,AP 值进一步提高到 35.4%;本文网络包含 3 个子模块 G-ResNet50, MSA 机制和 CIoU,叠加后的 AP 值达到了 35.7%,优于任何单个模块的性能,证明该网络在特征提取和重要特征聚焦方面取得了很好的效果.

表 7 添加不同子模块时网络的 AP 对比 %

子模块	AP
ResNet50+FPN+GIoU(VFNet)	34.1
G-ResNet50+FPN+GIoU	34.7
ResNet50+FPN+MSA+GIoU	34.8
G-ResNet50+FPN+MSA+GIoU	35.4
G-ResNet50+FPN+MSA+CIoU(本文)	35.7

类似地,在 COCO 2017 数据集上进行特征可视化操作,结果如图 6 所示.可以看出,图 6b 和图 6d 中,在未添加 MSA 机制的情况下,网络对图像的注意力是分散的,从原始主干网络中提取的对象特征权重并不高;而添加 MSA 机制后,网络的注意力集中在目标对象上,上下文特征提取中的信息会被聚合,重要的信息会被赋予更高的权

重(如图 6 中的亮点). 实验结果表明, MSA 机制可以使算法更加关注于感兴趣的区域.

将本文网络与其他经典网络进行实验, 包括 Faster R-CNN^[6], Cascade R-CNN^[7], RetinaNet^[11], FCOS^[12], ATSS^[15], FoveaBox^[16], MFANet^[23], SSD512^[40], PAA^[41]和 IQDet^[42], 这些网络的提出时间具有较大的跨度, 充分体现了目标检测的不

断发展过程. 实验结果如表 8 所示. 可以看出, SSD512 的 AP 值只有 25.6%, 这是因为该网络采用 VGG16 作为主干网络, 所以检测精度没有达到很高的水平; Faster R-CNN 也未能取得很好的结果, AP 值仅为 28.3%; PAA 和 Cascade R-CNN 得到了 34.2%和 34.7%的高 AP 值; 本文网络的 AP 值为 35.7%, 高于对比网络, 同时, 其 AP_{75} , AP_S , AP_M

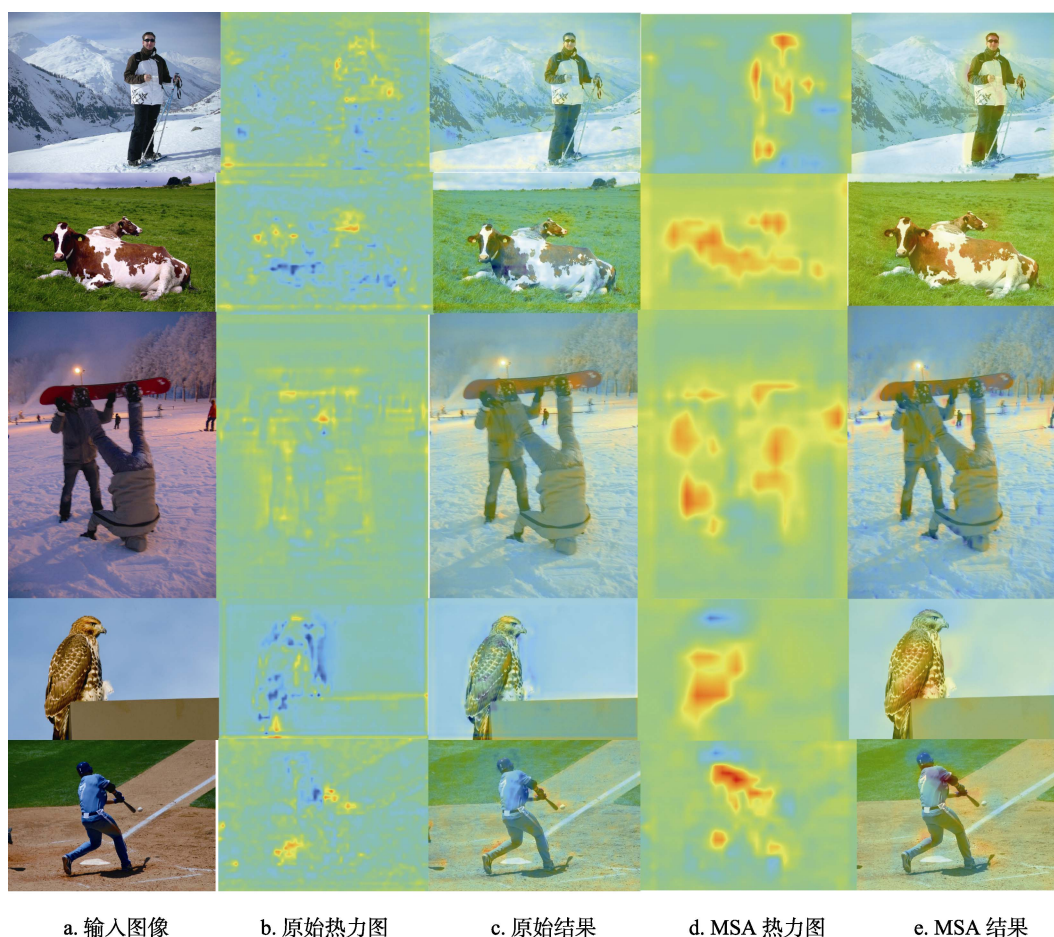


图 6 COCO 2017 数据集上的可视化结果

表 8 11 种网络实验结果比较

网络	主干网络	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SSD512 ^[40]	VGG16	25.6	44.0	26.2	9.1	29.0	38.9
Faster R-CNN ^[6]	ResNet50	28.3	45.0	30.4	12.0	31.0	41.7
FCOS ^[12]	ResNet50	28.7	45.0	29.9	9.6	31.9	44.3
FoveaBox ^[16]	ResNet50	29.5	47.4	31.0	10.9	33.0	45.8
RetinaNet ^[11]	ResNet50	30.5	47.3	32.2	10.8	35.2	48.4
ATSS ^[15]	ResNet50	32.7	48.7	34.8	13.5	36.8	49.8
PAA ^[41]	ResNet50	34.2	49.8	37.0	14.3	38.3	52.4
Cascade R-CNN ^[7]	ResNet50	34.7	52.1	37.3	13.6	38.1	52.4
IQDet ^[42]	ResNet50	33.7	50.0	36.1	13.9	37.8	50.0
MFANet ^[23]	ResNet50	34.2	50.6	36.6	16.1	38.2	50.3
本文	G-ResNet50	35.7	51.4	38.5	17.4	39.8	50.7

注. 粗体表示最优值.

也是所有网络中最高的,有效地平衡了网络对不同大小物体的检测效果,证明了本文网络的可行性。

目标检测算法非常重要的指标之一就是检测速度,在相同的实验设备上,将本文网络与其他网络的检测速度进行实验,结果如表 9 所示。可以看出,本文网络的目标检测速度是 34.1 幅/s,排第 5 名,但其检测速度只比排第 1 名的 YOLO V5 网络慢 9.5 幅/s;本文网络在目标检测精度方面表现优异,AP 达到 35.7%,高于其他网络。实验结果表明,目标检测算法往往难以在速度和精度之间取得平衡,除了本文网络和 YOLO V5 网络以外,其他检测速度较快的算法在精度方面大多表现不佳,如 FoveaBox 在取得 42.8 幅/s 的速度同时,AP 只有 29.7%;而精度较好的同时又容易造成速度较低,如 PAA 的 AP 为 34.2%,但检测速度仅为 5.4 幅/s;相比之下,本文网络在检测速度和精度方面得到较好的兼顾,很好地完成了目标检测任务。

本文网络的检测效果结果如图 7 所示。可以看出,图像③中只有一只鸟时,本文网络可以准确地检测到物体;有多个目标时,该也能达到很好的检测效果,如图像⑤和图像⑨等;对于图像①和图像②,即使一些目标出现被遮挡的情况,网络仍然可

表 9 10 种网络检测速度对比

网络	速度		精度	
	FPS/(幅·s ⁻¹)	排名	AP/%	排名
SSD512 ^[40]	23.2	9	25.6	10
Faster R-CNN ^[6]	29.6	8	28.3	9
FCOS ^[12]	41.0	3	28.7	8
FoveaBox ^[16]	42.8	2	29.5	7
RetinaNet ^[11]	31.8	7	30.5	6
ATSS ^[15]	33.9	6	32.7	5
PAA ^[41]	5.4	10	34.2	2
IQDet ^[42]	38.9	4	33.7	4
YOLO V5	43.6	1	33.9	3
本文	34.1	5	35.7	1

以准确地识别其类别;此外,本文网络很好地完成了对小目标和模糊对象的检测任务,如图像⑦中的人和图像⑧汽车中的人;但是,由于数据集的种类有限,同时图像的数量庞大,个别对象也会出现被错误检测的情况,如图像⑥中的圆形野营帐篷被检测为雨伞,这些错误检测较为少见,在误差允许范围内不会影响整体检测结果。实验结果表明,本文网络完成了精准的目标检测的任务,具有出色的识别效果。



图 7 本文网络检测效果

4 结 语

本文提出一种基于分组残差及变换特征维度的目标检测网络. 该网络使用新型主干结构替换原始网络结构, 同时添加 MSA 机制, 有效地提取不同尺度目标的特征, 增强上下文信息的聚合, 提高了检测的准确性. 在相同的配置和平台下, 本文网络的 AP, AP₅₀, AP₇₅, AP_S, AP_M 和 AP_L 分别提高 1.6 个百分点、1.7 个百分点、2.0 个百分点、3.0 个百分点、2.2 个百分点和 0.2 个百分点. 未来将研究特征融合在通道维度和空间维度上的差异, 以及探讨它们各自对不同尺寸物体检测精度的影响.

参考文献(References):

- [1] Luo Shun, Yu Juan. Pedestrian target detection algorithm based on improved multi-scale network[J]. Journal of Fuzhou University (Natural Science Edition), 2022, 50(5): 587-594(in Chinese)
(罗舜, 于娟. 改进多尺度网络的行人目标检测算法[J]. 福州大学学报(自然科学版), 2022, 50(5): 587-594)
- [2] Han Shuai, Luo Suyun, Chen Yangzhong. Vehicle detection algorithm based on improved YOLOv4 model[J]. Intelligent Computer and Applications, 2022, 12(2): 115-119(in Chinese)
(韩帅, 罗素云, 陈杨钟. 改进 YOLOv4 模型的车辆检测算法[J]. 智能计算机与应用, 2022, 12(2): 115-119)
- [3] Chao Yuan, Liu Wenhui, Tang Hanbing, et al. Fast indoor face detection method based on improved YOLO-v4[J]. Computer Engineering and Applications, 2022, 58(14): 105-113(in Chinese)
(巢渊, 刘文汇, 唐寒冰, 等. 基于改进 YOLO-v4 的室内人脸快速检测方法[J]. 计算机工程与应用, 2022, 58(14): 105-113)
- [4] Chen Daqian, Zhang Fan, Hao Pengyi, et al. 2D medical image segmentation combining multi-scale channel attention and boundary enhancement[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(11): 1742-1752(in Chinese)
(陈大千, 张凡, 郝鹏翼, 等. 结合多尺度通道注意力和边界增强的 2D 医学图像分割[J]. 计算机辅助设计与图形学学报, 2022, 34(11): 1742-1752)
- [5] Sha Miaomiao, Li Yu, Li An. Multiscale aircraft detection in optical remote sensing imagery based on advanced Faster R-CNN[J]. National Remote Sensing Bulletin, 2022, 26(8): 1624-1635(in Chinese)
(沙苗苗, 李宇, 李安. 改进 Faster R-CNN 的遥感图像多尺度飞机目标检测[J]. 遥感学报, 2022, 26(8): 1624-1635)
- [6] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C] // Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 91-99
- [7] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6154-6162
- [8] Pang J M, Chen K, Shi J P, et al. Libra R-CNN: towards balanced learning for object detection[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 821-830
- [9] Duan K W, Xie L X, Qi H G, et al. Corner proposal network for anchor-free, two-stage object detection[C] // Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 399-416
- [10] Zhou X Y, Koltun V, Krähenbühl P. Probabilistic two-stage detection[OL]. [2022-12-02]. <https://arxiv.org/abs/2103.07461>
- [11] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C] // Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2999-3007
- [12] Tian Z, Shen C H, Chen H, et al. FCOS: fully convolutional one-stage object detection[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 9626-9635
- [13] Dong Z W, Li G X, Liao Y, et al. CentripetalNet: pursuing high-quality keypoint pairs for object detection[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10516-10525
- [14] Dai X Y, Chen Y P, Xiao B, et al. Dynamic head: unifying object detection heads with attentions[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 7369-7378
- [15] Zhang S F, Chi C, Yao Y Q, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9756-9765
- [16] Kong T, Sun F C, Liu H P, et al. FoveaBox: beyond anchor-based object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 7389-7398
- [17] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2022-12-02]. <https://www.emanticscholar.org/reader/eb42cf88027de515750f230b23b1a057dc782108>
- [18] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1-9
- [19] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
- [20] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2261-2269
- [21] Redmon J, Farhadi A. YOLOv3: an incremental improvement[OL]. [2022-12-02]. <https://arxiv.org/abs/1804.02767>
- [22] Zhang H, Wu C R, Zhang Z Y, et al. ResNeSt: split-attention

- networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2022: 2735-2745
- [23] Wang G H, Gan X, Cao Q C, *et al.* MFANet: multi-scale feature fusion network with attention mechanism[J]. The Visual Computer, 2023, 39(7): 2969-2980
- [24] Yao Tingting, Zhang Bo, Liu Xiaoming. Feature enhanced fully convolutional network for ship detection[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(7): 1028-1036(in Chinese)
(姚婷婷, 张波, 柳晓鸣. 特征增强全卷积网络下的船舶检测[J]. 计算机辅助设计与图形学学报, 2022, 34(7): 1028-1036)
- [25] Wang G H, Zhai Q Y. Feature fusion network based on strip pooling[J]. Scientific Reports, 2021, 11(1): Article No.21270
- [26] Wang G H, Zhai Q Y, Lin J H. Multi-scale network for remote sensing segmentation[J]. IET Image Processing, 2022, 16(6): 1742-1751
- [27] Wang G H, Zhai Q Y, Liu H. Cross self-attention network for 3D point cloud[J]. Knowledge-Based Systems, 2022, 247: Article No.108769
- [28] Jin Yan, Xue Zhizhong, Jiang Zhiwei. Medical image segmentation based on recurrent residual convolution neural network[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(8): 1205-1215(in Chinese)
(金燕, 薛智中, 姜智伟. 基于循环残差卷积神经网络的医学图像分割算法[J]. 计算机辅助设计与图形学学报, 2022, 34(8): 1205-1215)
- [29] Wang G H, Lin J H, Zhai Q Y, *et al.* MAMask: multi-feature aggregation instance segmentation with pyramid attention mechanism[J]. IET Image Processing, 2022, 16(5): 1341-1348
- [30] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141
- [31] Li X, Wang W H, Hu X L, *et al.* Selective kernel networks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 510-519
- [32] Woo S, Park J, Lee J Y, *et al.* CBAM: convolutional block attention module[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19
- [33] Fu J, Liu J, Tian H J, *et al.* Dual attention network for scene segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3141-3149
- [34] Jiang X H, Zhang L, Xu M L, *et al.* Attention scaling for crowd counting[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 4705-4714
- [35] Zhang H Y, Wang Y, Dayoub F, *et al.* VarifocalNet: an IoU-aware dense object detector[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 8510-8519
- [36] Rezatofighi H, Tsoi N, Gwak J, *et al.* Generalized intersection over union: a metric and a loss for bounding box regression[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 658-666
- [37] Zheng Z H, Wang P, Liu W, *et al.* Distance-IoU loss: faster and better learning for bounding box regression[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 12993-13000
- [38] Gao Z L, Xie J T, Wang Q L, *et al.* Global second-order pooling convolutional networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3019-3028
- [39] Wang Q L, Wu B G, Zhu P F, *et al.* ECA-Net: efficient channel attention for deep convolutional neural networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 11531-11539
- [40] Liu W, Anguelov D, Erhan D, *et al.* SSD: single shot MultiBox detector[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 21-37
- [41] Kim K, Lee H S. Probabilistic anchor assignment with IoU prediction for object detection[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 355-371
- [42] Ma Y C, Liu S T, Li Z M, *et al.* IQDet: instance-wise quality distribution sampling for object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 1717-1725